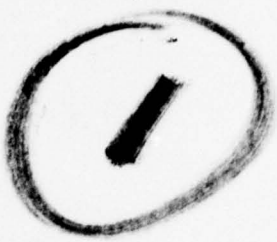
LEVEL II



Research Memorandum 78-17

AD A077965

ANALYSIS OF VARIANCE: SELECTION OF A MODEL AND SUMMARY STATISTICS

Frederick H. Steinheiser, Jr., and Kenneth I. Epstein

UNIT TRAINING & EVALUATION SYSTEMS TECHNICAL AREA

U. S. Army
Research Institute for the Behavioral and Social Sciences

August 1978

Army Project Number
2Q762722A764

Unit Training
Evaluation

12

9 Research memo.,

Research Memorandum 78-17

6 ANALYSIS OF VARIANCE: SELECTION OF A MODEL
AND SUMMARY STATISTICS

10 Frederick H. Steinheiser, Jr., Kenneth I. Epstein

14 ARI-RM-78-17

Submitted by:
Frank J. Harris, Chief
UNIT TRAINING AND EVALUATION SYSTEMS TECHNICAL AREA

11 August 1978

Approved by:

A. H. Birnbaum
Acting Director
Organizations and Systems Research
Laboratory

Joseph Zeidner, Technical Director
(Designate)
U.S. Army Research Institute for
the Behavioral and Social Sciences

Research Memorandums are informal reports on technical research problems. Limited distribution is made, primarily to personnel engaged in research for the Army Research Institute.

ANALYSIS OF VARIANCE: SELECTION OF A MODEL AND SUMMARY STATISTICS

The topics of this paper are models for the analysis of variance (ANOVA) (fixed, random, or mixed models) and the subsequent summary statistics (F ratio, quasi-F ratio, and magnitude of treatment effect) that may be computed following the ANOVA. ANOVA is a useful method for assessing the statistical significance of treatment effects. But the significance of an effect is a function of two decisions. The first decision is the selection of a model and an appropriate sampling plan for elements within each of the treatment factors. The second decision is the choice of summary statistics that indicate the extent of significance achieved. In this paper, comparisons will be made between models and between summary statistics. Specific issues will be clarified concerning the interpretation of results when various models and summary statistics are used on the same set of data.

Selection of an ANOVA Model

In the fixed-effects model, the levels of the independent variables are assumed to have been exhaustively sampled. No generalization beyond those levels sampled is intended or theoretically permissible. The random effects model assumes that the selected treatment variables have been randomly selected from a very large population of such variables. Generalization of results from the random sample to the population is allowed. The mixed model allows both fixed and random factors to be studied in the same experiment, and the results for each factor are to be interpreted according to that factor's sampling plan.

The choice of a model has an impact on the probability of obtaining the observations under the null hypothesis for each treatment (factor). Behavioral research is particularly vulnerable to the choice of a model because often the investigator can use only a limited sample of the possible number of stimuli (items, drug doses, etc.). Furthermore, the same stimulus set may, by necessity, be given to all subjects because of the difficulty in creating comparable sets of stimuli.

As a simple hypothetical experiment (adapted from Clark, 1973), suppose that two classes of stimuli, nouns and verbs, are individually shown to subjects. The purpose is to see if the subject takes the same amount of time to identify each word as a member of the correct

---

Presented at the 23rd Conference on the Design of Experiments in Army Research, Development, and Testing; Naval Postgraduate School, Monterey, Calif.: 19 October 1977.

part-of-speech class. This simple hypothesis will be shown to have interesting implications for both experimental design and statistical analysis.

First, fixed sets of nouns and verbs that are matched on relevant parameters, such as number of letters and frequency of occurrence, are prepared. To generalize to the full domain of nouns and verbs, each subject should receive a <u>different</u> random sample of words from the two lists. It is impossible, however, to match the words on <u>all</u> relevant variables. It is also practically impossible to use a different random sample of words for each subject.

Consider the experimental design shown in Table 1, in which "s" subjects each are presented "w" different nouns and verbs. To compare the adequacy of the several possible F ratios for testing the difference in response time to the two "treatment" (part of speech) conditions, Table 2 and Table 3, which show expected mean squares (EMS), will be helpful.

Table 1

Assignment of Subjects and Parts of Speech

Subject	Part of speech	
	P_1 (nouns)	P_p (verbs)
S_1	$W_1 \cdots W_{w/2}$	$W_{w/2+1} \cdots W_w$
..		
..		
S_s		

If the significance of the Parts of Speech treatment is tested, the appropriate F ratio for the model illustrated in Table 2 is $F_1 = MS_p/MS_{pxs}$. The only term in the numerator that is not in the denominator is $sw\sigma^2_p$. However, if this same F ratio is used with the model in Table 3 (applicable when generalization is desired to all nouns and verbs), then this F ratio will contain two terms that are not in the denominator: $so^2_{w(p)}$ and $sw\sigma^2_p$. Using alternative error terms in the parts-of-speech fixed, words-random model (Table 2) also leads to the same problem. For example, if we test the parts-of-speech effect against the words within parts-of-speech effect, we obtain $F_2 = MS_p/MS_{w(p)}$. In this case, EMS_p exceeds $EMS_{w(p)}$ by the amount of

Table 2

(EMS), Assuming Parts of Speech Is a Fixed Factor and Subjects and Words Are Random

Source	EMS
P (Part of speech)	$\sigma_e^2 + sw\sigma_p^2 + s\sigma_{w(p)}^2 + w\sigma_{pxs}^2 + \sigma_{sxw(p)}^2$
W(P) (Words within part of speech)	$\sigma_e^2 + s\sigma_{w(p)}^2 + \sigma_{sxw(p)}^2$
S (Subjects)	$\sigma_e^2 + pw\sigma_s^2 + \sigma_{sxw(p)}^2$
P x S	$\sigma_e^2 + w\sigma_{pxs}^2 + \sigma_{sxw(p)}^2$
S x W(P)	$\sigma_e^2 + \sigma_{sxw(p)}^2$

Table 3

(EMS), Assuming Parts of Speech and Words Are Fixed and Subjects Are Random

Source	EMS
P	$\sigma_e^2 + sw\sigma_p^2 + w\sigma_{pxs}^2$
W(P)	$\sigma_e^2 + s\sigma_{w(p)}^2 + \sigma_{sxw(p)}^2$
S	$\sigma_e^2 + pw\sigma_s^2$
P x S	$\sigma_e^2 + w\sigma_{pxs}^2$
S x W(P)	$\sigma_e^2 + \sigma_{sxw(p)}^2$

$w\sigma^2_{pxs} + ws\sigma^2_p$. Therefore, this F_2 ratio would also be significant when the true contribution of σ^2_p due to parts of speech (treatments) is really zero. In summary, both F_1 and F_2 could be significant when $\sigma^2_p = 0$, provided that σ^2_w and σ^2_{pxs} exceed zero.

A possible solution to this dilemma is to take the "quasi-F" ratio, or F', which equals $(MS_p + MS_{sw(p)})/(MS_{pxs} + MS_{w(p)})$. Now the only term in the numerator that is not in the denominator is σ^2_p. However, F' is distributed only approximately as F, although the error involved is not large provided that adjustments are made to the degrees of freedom.

A more conservative solution is minimum F', which assumes that $MS_{sxw(p)}$ is zero. A detailed discussion of this problem may be found in Clark (1973).

A series of Monte Carlo computer simulations (Forster & Dickinson, 1976) explored the relationship between all of the above F ratios and the resulting Type I error rates. Generally, F_1 and F_2 alone produced unacceptably high error rates, whereas F' and min F' were more conservative, as shown in Table 4.

As shown in Table 5, increasing the number of items and subjects tends to decrease F_1 Type I error for the fixed effects model where only subjects are random. Min F' and F' continue to have lower error rates.

The "Magnitude of Effect" as a Summary Statistic

The F ratio indicates the level of statistical significance that can be attributed to a particular treatment. The degree of statistical significance is a joint function of the "true" strength of that factor, the error variability, which reflects the degree of experimental control, and the sample size (i.e., number of subjects tested). As sample size increases, there is increasing power to reject a false null hypothesis. Thus, in conducting large-scale experiments with hundreds of subjects, the large "n" may be necessary to detect a weak "signal" buried in a background of "noisy" data. But the large n may also lead to spuriously significant F ratios that are actually statistical artifacts.

One index for assessing the significance of effects is the "magnitude of effect," also referred to as the "proportion of variance accounted for." It is interesting to note that relatively few research papers have included this index as compared to the occurrence of ubiquitous F ratio. Basically, the magnitude of effect (m.e.) measures the degree of association between the independent variable(s) and the

Table 4

Type I Error Rates as a Function of Variation in MS_{sxp} and $MS_{w(p)}$

Source of variance manipulated	$S.D._1$	$S.D._2$	F_1	F_2	min F'	F'
Neither	0	0	.044	.046	.010	.026
$MS_{w(p)}$	5	0	.228	.052	.038	.044
	10	0	.484	.070	.060	.060
	15	0	.586	.056	.048	.052
	20	0	.724	.050	.048	.048
MS_{sxp}	0	5	.042	.146	.024	.036
	0	10	.064	.388	.042	.042
	0	15	.036	.520	.032	.034
	0	20	.042	.588	.038	.042
Both	5	5	.124	.096	.034	.042
	10	10	.190	.090	.040	.040
	15	15	.220	.138	.056	.064
	20	20	.208	.118	.048	.048

Note: 500 observations per situation, alpha = .05, p = 2, q = 5, r = 9.

Table 5

Type I Error Rates as a Function of the Numbers of Subjects and Items

Subjects	Items	F_1	F_2	min F'	F'
10	5	.240	.070	.040	.040
10	20	.090	.290	.053	.053
20	5	.307	.077	.067	.067
20	20	.193	.217	.060	.060

Note: 300 observations per situation, $S.D._1 = S.D._2 = 20$, and alpha = .05.

dependent variable(s). In the simplest case for ANOVA having fixed factors, none of which are repeated, the m.e. formula is magnitude of effect = $(SS_{effect} - df_{effect} \times MS_{error})/(SS_{total} + MS_{error})$. Rules for deriving m.e. indexes are provided by Dodd and Schultz (1973), along with tables for representative ANOVA designs.

The present paper is concerned with the interpretation of these summary statistics, because both F and m.e. can be computed from the same set of data. It is clear that as the statistical significance for a given effect increases--that is, the p(observation/null) decreases--the magnitude for that effect also increases. But it is also possible that an F ratio may be highly statistically significant, yet the m.e. for that effect could account only for a very small proportion of the overall variance. The results from an experiment summarized in the following section show that when statistical significance ($p < .001$) was achieved by several treatments, the m.e. for these treatments ranged from 1% to 23%.

A Study of Marksmanship

An experiment was conducted for the U.S. Army Military Police School at Fort McClellan, Ala., in which 237 students each shot a total of 240 handgun rounds from eight different position-distance combinations. There were three repetitions of 80 shots each, at stationary silhouette targets. Within each repetition, 5 shots were taken, the weapon was reloaded, and 5 more shots were fired in the adjacent test lane. (Each subject had previously passed a training course with a score of at least 35 hits out of 50 shots.) In the test, 160 trials (two repetitions) were taken on Thursdays and the third repetition was taken on Fridays. The completely crossed design was therefore A x B x C x D, or 237 x 2 x 8 x 3, or subjects x lanes x tables x repetitions.

Table 6 highlights the results of the ANOVA from this experiment. The first column of F ratios assumes a mixed model, with B, C, D as fixed factors. The second column of F ratios assumes that only the Tables factor was a fixed factor. The third F ratio column assumes that all four factors were randomly sampled from their respective populations. The point is rather obvious: Different ANOVA models produce different F ratios for null hypothesis rejection, given the same set of data.

The problem of interpreting the F ratios needs to be addressed. Is there, for example, a significant effect due to Lanes or to Repetitions? If these effects are assumed to be fixed, the answer is yes; if they are assumed to be random, the answer for Lanes is no; and for repetitions the level of statistical significance has decreased greatly.

Table 6

Changes in F Ratios as a Function of ANOVA Model

Source	df[a]	MS	F[b]	F[c]	F[d]
A (Subjects)	236	12.80		3.93****	2.54****
B (Lanes)	1	7.70	7.33****	5.96**	2.26
C (Tables)	7	732.71	385.64****	79.11****	79.11***
D (Repetitions)	2	34.75	14.18****	12.55****	4.71**

****p < .001.
***p < .01.
**p < .025.

[a]df for F ratios were obtained using the Satterthwaite approximation.

[b]A random; B, C, D fixed effects.

[c]A, B, D random; C fixed effects.

[d]A, B, C, D all random effects.

We offer the suggestion that the choice of the ANOVA model, and ultimately the level of significance reached, lies in the eyes of the beholder--the scientist. From a sponsor's perspective, only those conditions studied in the experiment may be of interest. If many lanes, repetitions, or even tables are never to be studied or added to the sponsor's testing program, then those factors would never be sampled from a larger population of such factors. One might argue from a scientific point of view, however, that many additional lanes, repetitions, and firing positions could have been tested; that is, we happen to have chosen only three repetitions, two lanes per subject, and eight different distance-position combinations. Thus, the sponsor-practitioner wants information that is specific to his or her particular test. In contrast, the scientific "purist" may perceive this one test or experiment as merely one of many different kinds that could have been conducted for the sponsor. Hence, the choice of model indeed influences the significance levels obtained.

The power of the F ratio to reject a false null hypothesis is a function of (a) the "true" strength of the particular factor and (b) the sample size. Although a large sample size may help to detect a weak signal in a noisy background, using such a large sample can lead to increasingly significant F ratios with little, if any, concomitant increase in the m.e. It is to this latter summary statistic that we now turn our attention, in the analysis of the same set of marksmanship data.

The m.e. results are shown in Table 7. This table shows that the largest effect other than random error was due to the Tables factor, which captured a 23% share of the total score variability. The effect due to the Subject factor, which reflected individual differences among the students, reached nearly 10%. Several interaction terms, in which Tables was a factor, accounted for about 6% to 7%.

Table 7

Changes in Magnitude of Effect Index as a Function of ANOVA Model

	Proportion of total variance		
Source	A random, B,C,D fixed	A,B,D random, C fixed	A,B,C,D random
A (Subjects)	.0852	.1027	.1030
B (Lanes)	.0004	.0006	.0005
C (Tables)	.1643	.2454	.2631
D (Repetitions)	.0027	.0041	.0042

Note that the effect due to Repetitions in Table 5 was statistically significant, whereas Repetitions contributed an effect worth only about .4% in Table 7. The reason for this apparent discrepancy between the two summary statistics is due to the large number of subjects, which in turn produced a large number of degrees of freedom. This allows small F ratios to achieve statistical significance more readily. Thus, the values for m.e. in Table 7 act as a check upon the significance levels listed in Table 6. Therefore, the effect due to Repetitions reveals a slight, but probably inconsequential, learning effect. A similar line of reasoning holds for the interpretation of the Lanes variable in Tables 6 and 7.

Summary and Conclusions

In actual experimental testing situations, it may not be easy to determine whether a given treatment should be classified as a fixed or as a random effect. For example, in the experiment outlined, the Lanes, Repetitions, and Tables factors could be considered as either fixed or random. The Tables factor had eight levels, representing the eight specific position-distance combinations that comprise the marksmanship test. Because there are theoretically an infinite number of distance-position combinations, Tables could be interpreted as a sampling of eight from this much larger population. A random effects assignment to Tables could easily be justified because an experimenter is often interested in generalizing results beyond the specific treatment levels to a larger set of "real-world" circumstances. Furthermore, the probability of falsely rejecting a true null hypothesis is less when a treatment is considered to be random.

In sum, the wise use of an ANOVA model involves (1) determination of fixed versus random factors, (2) computation of complete sets of summary statistics, and (3) interpretation of the statistics.

REFERENCES

Clark, H. H. The Language-as-Fixed-Effect-Fallacy: A Critique of Language Statistics in Psychological Research. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 335-359.

Dodd, D. H., & Schultz, R. F. Computational Procedures for Estimating Magnitude of Effect for Some Analysis of Variance Designs. Psychological Bulletin, 1973, 79, 392-395.

Forster, K. I., & Dickinson, R. G. More on the Language-as-Fixed-Effect Fallacy: Monte Carlo Estimates of Error Rates for F_1, F_2, F' and min F'. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 135-142.